AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 132, No. 6
Printed in U.S.A.

THE IMPACT OF DIETARY MEASUREMENT ERROR ON PLANNING SAMPLE SIZE REQUIRED IN A COHORT STUDY

LAURENCE S. FREEDMAN,[1] ARTHUR SCHATZKIN,[2] AND YOHANAN WAX[1]

Freedman, L. S. (National Cancer Inst., NIH, Bethesda, MD 20892), A. Schatzkin, and Y. Wax. The impact of dietary measurement error on planning sample size required in a cohort study. *Am J Epidemiol* 1990;132:1185–95.

Dietary measurement error has two consequences relevant to epidemiologic studies: first, a proportion of subjects are misclassified into the wrong groups, and second, the distribution of reported intakes is wider than the distribution of true intakes. While the first effect has been dealt with by several other authors, the second effect has not received as much attention. Using a simple errors-in-measurement model, the authors investigate the implications of measurement error for the distribution of fat intake. They then show how the inference of a more narrow distribution of true intakes affects the calculation of sample size for a cohort study. The authors give an example of the calculation for a cohort study investigating dietary fat and colorectal cancer. This shows that measurement error has a profound effect on sample size, requiring a six- to eightfold increase over the number required in the absence of error, if the correlation coefficient between reported and true intakes is 0.65. Reliable detection of a relative risk of 1.36 between a true intake of greater than 47.5% calories from fat and less than 25% calories from fat would require approximately one million subjects.

cohort studies; colorectal neoplasms; dietary fats

It is possible that dietary factors are a contributing cause of many chronic diseases, and a considerable research effort is being spent investigating several plausible dietary hypotheses. Epidemiologic studies of diet and disease are made more difficult by the problem of dietary measurement. Although progress has been made in the development of dietary questionnaires (1), comparison of questionnaires with supposedly more accurate food records shows relatively poor agreement (2). This indicates a considerable discrepancy between the true diet of an individual and that inferred from a questionnaire, although it would be wrong to assume without question that the food record itself comes close to the truth. If, as is usual, the aim is to measure average intake over the long term, the questionnaire may be in error due to failure of recall, whereas a food record over a short period may fail to accurately represent long-term diet. Most likely, therefore, both the questionnaire *and* the food record are in error. However, the mean of several food records taken over a broad spectrum of time is likely to perform better than a single food questionnaire.

Several authors (3–7) have recently writ-

Received for publication September 12, 1989, and in final form June 29, 1990.

[1] Biometry Branch, Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, MD.

[2] Cancer Prevention Studies Branch, Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, MD.

Reprint requests to Laurence S. Freedman, Biometry Branch, DCPC, National Cancer Institute, Executive Plaza North, Suite 344, Bethesda, MD 20892.

The authors wish to thank Dr. Gladys Block for useful discussion and access to her analysis of the Women's Health Trial Data and Anne Hartman for valuable advice.

ten about the effect of measurement error on the power of a cohort or case-control study to detect a relation between a dietary component and a disease. All have commented on the loss of power due to dietary measurement error, and some express consequent doubt as to whether a negative (i.e., nonsignificant) result in a conventionally sized study should be taken as evidence of the nonimportance of a particular effect (3, 5).

In this paper, we consider the required size of a cohort study designed to detect an effect of a dietary variable on the incidence of a cancer, assuming the dietary variable is imprecisely measured. One consequence of imprecision is that subjects are misclassified into the wrong exposure groups. Walker and Blettner (7), in a seminal paper on the effects of such misclassification in epidemiologic studies, tabulated required sample sizes for hypothetical examples of a cohort and a case-control study. A second consequence of dietary measurement error is that the distribution of true intakes of a nutrient is narrower than the distribution of reported intakes, i.e., the true intakes are more homogeneous among the population than are those reported. Prentice et al. (8) noted that this effect reduced the power of a cohort study to detect given effects of nutrient intake, in their discussion of dietary fat and breast cancer. We will discuss this "narrowing" phenomenon in more detail and incorporate its effect together with that of misclassification into our sample size calculations.

Distributions of reported and true dietary intakes

When planning the sample size of a cohort study, it is necessary to consider the distribution of intakes of specific nutrients in the population. We focus on dietary fat as a nutrient of possible importance in the etiology of cancer (9) and on the distribution of its intake among a cohort of men and women over age 50 years.

Suppose we measure dietary fat intake by the percent of calories from fat consumed by an individual. At the planning stage, it is probable that no direct information on the distribution of intakes in our cohort is available. However, we suppose that data from the National Health Interview Survey, conducted in 1987 (10), are the best available for our purpose. These indicate that in subjects over age 50 years, percent calories from fat has a distribution shape very close to the normal, with a mean of 37.7 and a standard deviation of 7.6, and that 16 percent of subjects report eating less than 30 percent calories from fat and 5 percent report eating less than 25 percent calories from fat (table 1). However, these reported intakes are obtained from a food questionnaire (similar to the instrument we plan to use in our study) and are subject to errors. What can we say about the distribution of true intakes?

To answer this question, we need to specify a statistical model describing the nature of the errors made in reporting percent calories from fat intake. Let X denote the true percent calories from fat and Y denote reported percent calories from fat. Then we assume that

$$Y = X + \epsilon, \qquad (1)$$

i.e., that the reported intake, Y, is the sum of the true intake, X, and some independent error, ϵ, which may be positive or negative. We further assume that the average ϵ is 0

Table 1

Distribution of reported and true intakes of fat, measured by percent calories from fat

	Reported percent calories from fat	True percent calories from fat*
Mean	37.7	37.7
Standard deviation	7.6	4.9
Percent consuming less than		
25%	5	0.5
30%	16	6
35%	36	29
40%	62	68
45%	83	93
50%	95	99.4

* Calculated assuming model 1 and $\rho = 0.65$.

of calories from fat con-
dividual. At the planning
able that no direct infor-
stribution of intakes in our
ble. However, we suppose
he National Health Inter-
nducted in 1987 (10), are
le for our purpose. These
subjects over age 50 years,
from fat has a distribution
o the normal, with a mean
ndard deviation of 7.6, and
of subjects report eating
ent calories from fat and
eating less than 25 percent
(table 1). However, these
are obtained from a food
nilar to the instrument we
study) and are subject to
we say about the distri-
akes?

question, we need to spec-
odel describing the nature
ade in reporting percent
intake. Let X denote the
ies from fat and Y denote
calories from fat. Then we

$$= X + \epsilon, \qquad (1)$$

ted intake, Y, is the sum
X, and some independent
y be positive or negative.
e that the average ϵ is 0

TABLE 1

orted and true intakes of fat, ercent calories from fat

eported percent alories from fat	True percent calories from fat*
37.7	37.7
7.6	4.9
5	0.5
16	6
36	29
62	68
83	93
95	99.4

g model 1 and $\rho = 0.65$.

(i.e., that the questionnaire is unbiased) and that ϵ has a distribution among the population which is normal with a variance which is the same regardless of the true level of fat intake. These assumptions represent the simplest realistic model available for dietary measurement error and have been adopted by other authors (11). In the Discussion, we consider evidence concerning the validity of the model.

Under this model, it may be shown that: 1) if Y is normally distributed, then X is normally distributed; 2) the mean of X is equal to the mean of Y; and 3) if the correlation between X and Y is ρ, then the standard deviation of X is ρ times the standard deviation of Y; alternatively, the ratio of standard deviations of X to Y equals the correlation coefficient ρ (see Appendix).

Applying these results, we conclude that the true percent calories from fat has a normal distribution with mean 37.7 (the same as that reported) and standard deviation 7.6ρ. Thus, if we know the value of the correlation ρ, then the distribution of the true percent calories from fat can be completely determined. Although ρ cannot be measured directly, a rough estimate of its value may be obtained from studies in which the food questionnaire is validated with a more accurate assessment of diet such as the mean of several 4-day food records.

One such study conducted as part of the pilot phase of the Women's Health Trial demonstrated a correlation of approximately 0.65 between percent calories from fat measured by a diet history questionnaire and by the mean of three 4-day diet records (12). Since the food record gives an imperfect measure of true average intake, the value of 0.65 is an attenuated estimate of the correlation between questionnaire and true intake, and therefore requires an upward adjustment (13). However, intraindividual correlation also biases the estimate of the correlation (13) and in this case requires an almost equal adjustment in the opposite direction. (Further details of these adjustments are available from the authors.) Thus, the value of 0.65 provides our best estimate of the correlation.

Applying this value of ρ to the National Health Interview Survey data, we estimate that the standard deviation of the true percent calories from fat in our cohort population is 4.9. We may now estimate that the proportion of the cohort who actually eat less than 25 percent calories from fat is not

$$5 \text{ percent } \left(= \Phi\left(\frac{25 - 37.7}{7.6}\right),\right.$$

$$\left.\text{where } \Phi \text{ is the normal integral}\right), \text{ but}$$

$$0.5 \text{ percent } \left(= \Phi\left(\frac{25 - 37.7}{4.9}\right)\right) \text{ (table 1).}$$

Thus, if the error model is correct and the assumed value of 0.65 for ρ is not too low, then we must conclude that the distribution of intake of fat (measured by percent calories from fat) is considerably narrower than that reported and that, in particular, only a miniscule proportion of the cohort is truly eating less than 25 percent calories from fat.

IDENTIFICATION OF A GROUP WITH ATYPICAL DIETARY INTAKE

A difficulty with cohort studies of diet and disease has been the narrow range of intakes among indigenous populations (6). The previous section shows that this problem is perhaps more acute than is sometimes realized, since measurement error accounts for a proportion of the variation in reported intakes, and true intakes are consequently more narrowly spread than those reported.

In the case of dietary fat and breast cancer, it has been speculated that one reason for some negative results in cohort studies is the inclusion in the cohort of only a very small number of women eating low amounts of fat, say less than 25 percent calories from fat; perhaps it is only in these women that one will see a generally reduced incidence of breast cancer? To test this hypothesis, one would need to identify spe-

cifically such a group of women and compare them with a group eating "average" amounts of fat. A natural way to identify such a group would be to choose women who report their intake of fat to be less than 25 percent calories from fat. The resulting group would include some women who eat more than 25 percent calories from fat because of the error in reporting, but one might hope that the group would not be too badly contaminated. Such hopes are ill-founded! Using model 1 and assuming $\rho = 0.65$, one may calculate the expected distribution of true intakes among the selected group (table 2, second row). Surprisingly, only 7 percent of those who report eating less than 25 percent calories from fat actually will do so; 58 percent will eat 25–32.5 percent calories from fat, and 35 percent will eat more than 32.5 percent calories from fat. Similarly, more than half of those reporting an intake of 25–32.5 percent calories from fat will actually eat more than 32.5 percent calories from fat.

Making the selection criterion more stringent (e.g., reducing the upper limit of reported fat intake to 20 percent calories from fat) does not remedy this problem (table 2, first row). Thus, in the presence of measurement error, identification of a subgroup of people who eat low amounts of fat is a very difficult task. Only by minimizing error can this problem be overcome.

Specification of relative risks

Suppose we wish to study a possible relation between dietary fat intake and the incidence of colorectal cancer in a prospective cohort. Naturally, an important question is the size of cohort which is necessary to detect reliably the existence of such a relation. To calculate the sample size, we have to specify the level of relative risk which we wish to detect. We must distinguish two different ways of specifying relative risk: 1) describing a relative risk gradient over the quantiles of the distribution of percent calories from fat, or 2) describing a relative risk gradient over the absolute levels of percent calories from fat. Walker and Blettner (7), when considering the effect of measurement error on sample size, choose the first method, assuming a relative risk of 3 between the fifth and first quintiles of the true intake distribution. However, when a cohort study is being designed, this form of relative risk specification has disadvantages. First, in scientific terms, the effect of intake on disease incidence relates directly to the actual level of intake and only indirectly to the quantile of the intake distribution. Epidemiologic studies are conducted in many different populations, each of which has its own intake distribution. Thus, a relative risk of 3 between the fifth and first quintiles may represent quite different magnitudes of effect, according to

TABLE 2

Distribution of true intakes of fat among groups selected according to their reported intake*

Reported percent calories from fat	True percent calories from fat				
	<25	25–32.5	32.5–40	40–47.5	>47.5
≤20	14	67	19	—†	—
≤25	7	58	34	1	—
25–32.5	1	32	60	7	—
32.5–40	—	11	66	23	—
40–47.5	—	2	47	48	3
>47.5	—	—	20	65	15

* Calculated assuming model 1 and $\rho = 0.65$; mean reported percent calories from fat = 37.7, standard deviation = 7.6. The calculations involved computer evaluation of the bivariate normal integral; they were performed using the IMSL subroutine, MDBNOR (14).

† Entries with dash indicate less than 1%.

the cohort population of interest. Second, the distribution of true intakes in the cohort population is unknown and needs to be inferred from a sampled distribution of reported intakes, just as we described in the earlier part of this paper. This inference depends crucially on the correlation between the true and reported intakes in the sample. Often there is uncertainty over the inference because of lack of knowledge of this correlation, and it is necessary to calculate sample size over a series of assumed values of the correlation. However, each assumption yields a different estimate of the true intake distribution, thus changing the relative risk gradient over the *quantiles* of the distribution, the very basis for the sample size calculation; this results in confusion.

The latter problem is not treated by Walker and Blettner (7). They assume that the relative risks over the quantiles of the true intake distribution are given a priori and that these risks do not depend on the correlation between true and reported intakes. Hence, their results relating to the required sample size are conditional on knowing the distribution of true intake at the outset. They show how these required sample sizes are inflated because of the misclassification of subjects induced by dietary measurement error. In this paper, we make a different assumption—that the relative risks over absolute values of the true intake can be specified (point 2 above) on the basis of international correlations of disease incidence and food disappearance data, for example, but that the true intake distribution is not known directly. We assume that the reported intake distribution of the cohort is already known through a previous or baseline survey. Furthermore, we also assume that the method of reporting diet used in that survey bears the same correlation with true intake as does the dietary measurement method to be used in the cohort study. This last assumption is not central to the sample size calculation—entirely different values of the correlation for the survey questionnaire and the cohort study questionnaire could be chosen—but it simplifies tabulation of the effect of measurement error. Moreover, this assumption will often be close to the truth. In the Appendix, we outline the statistical methods and show how to incorporate different values of the correlation for survey and cohort questionnaires. By taking this approach, we account for two factors resulting from dietary measurement error: 1) misclassification of subjects as described by Walker and Blettner, and 2) greater homogeneity of the true intake distribution. These factors independently decrease statistical power and increase sample size requirements, as we will demonstrate below.

To investigate the sample size requirements of our study, we considered three different sets of relative risks for developing colorectal cancer (table 3); we used the same groupings of percent calories from fat as in table 2. The three rows of table 3 represent three different relative risk gradients, which we call small, moderate, and large. The large gradient specifies a relative risk of greater than two between the extreme upper and lower groups. The small

TABLE 3

Three relative risk hypotheses according to the true absolute fat intake level, measured by percent calories from fat

Gradient	True percent calories from fat				
	<25.0	25.0–32.5	32.5–40.0	40.0–47.5	>47.5
Small	1.0	1.09	1.18	1.27	1.36
Moderate	1.0	1.16	1.32	1.48	1.64
Large	1.0	1.33	1.67	2.00	2.33

gradient specifies a relative risk of 1.36 between these same two groups and represents a level of risk which is low, but which would still have an impact on the public health of the nation because of the high incidence of the disease.

Calculation of sample size

The hypotheses in table 3 describe relative risks according to *true* fat intake. When analyzing the results of a cohort study, we relate *reported* fat intake to disease incidence. Usually subjects are grouped according to their reported intakes, and a statistical test is used to detect a trend over the ordered groups. Quite often, the groups chosen are the quintiles of the reported intake distribution, and a trend test is employed using the scores 1–5 for the five ordered quintiles. However, in keeping with our use of absolute values of percent calories from fat in specifying the relative risk, we define our risk groups for analysis using the same groupings as in table 3 applied to the reported intakes.

To calculate sample size, we need to determine the expected risks within these five reported percent calories from fat groups. These are calculated by applying the percentages from the appropriate row of table 2 to the hypothesized relative risks contained in the chosen row of table 3. For example, to determine the risk in the lowest reported fat intake group (reported calories from fat ≤25 percent) under the small gradient hypothesis, we calculate

$$[(7 \times 1.0) + (58 \times 1.09) + (34 \times 1.18) + (1 \times 1.27) + (0 \times 1.36)]/100 = 1.116.$$

For the group 25–32.5 percent calories from fat, the risk is

$$[(1 \times 1.0) + (32 \times 1.09) + (60 \times 1.18) + (7 \times 1.27) + (0 \times 1.36)]/100 = 1.156.$$

With the lowest fat intake group as the reference, the relative risk in the group ≤25 percent calories from fat is 1.0, and that in the group 25–32.5 percent calories from fat is 1.156/1.116 = 1.036. Relative risks in the five reported groups are shown in table 4 for each of the three risk hypotheses, still assuming that the correlation $\rho = 0.65$. Comparison of table 4 with table 3 shows the considerable attenuation in the relative risk gradient because of measurement error.

Having determined the relative risks across the five risk groups, we may now apply a standard sample size formula to find the required number of cases C to give a power, $1 - \beta$, of detecting a trend in risk to be significant at the 5 percent level. The formula we use is taken from Breslow and Day (15) and is in the Appendix.

The number of cases required in the cohort study depends strongly on which of the relative risk gradients is hypothesized (table 5). With no dietary measurement error, the number of cases required varies from 160 to 1,500, according to the hypothesis. When measurement error is accounted for, assuming a correlation of 0.65, these numbers are multiplied by approximately six to eight. This multiplicative factor is larger than the approximately threefold increase determined by Walker and Blettner (7), since we include the shrinkage of the true intake distribution in our calculation.

Table 4

Relative risks in groups defined by reported percent calories from fat according to three relative risk hypotheses*

Gradient	Reported percent calories from fat				
	<25.0	25.0–32.5	32.5–40.0	40.0–47.5	>47.5
Small	1.0	1.04	1.07	1.10	1.13
Moderate	1.0	1.06	1.11	1.16	1.22
Large	1.0	1.10	1.19	1.29	1.39

* Assuming correlation $\rho = 0.65$.

TABLE 5

For fat and colorectal cancer, numbers required to detect a significant effect at 5% level with 90% power using trend test over 5% calories from fat groups: <25%, 25–32.5%, 32.5–40.0%, 40.0–47.5%, and ≥47.5%

Relative risk*	Correlation between measured and true exposure				
	0.60	0.65	0.70	0.75	1.00†
Small					
Cases‡	14,500	10,400	7,700	5,800	1,500
Cohort§	1,455,000	1,044,000	773,000	582,000	151,000
Moderate					
Cases	5,600	4,000	2,900	2,200	530
Cohort	562,000	401,000	291,000	221,000	53,000
Large					
Cases	2,000	1,400	1,000	740	160
Cohort	201,000	141,000	100,000	74,000	16,000

* Over 5 groups (<25, 25–32.5, 32.5–40.0, 40.0–47.5, and ≥47.5). Small: 1.00, 1.09, 1.18, 1.27, and 1.36; moderate: 1.00, 1.16, 1.32, 1.48, and 1.64; large: 1.00, 1.33, 1.67, 2.00, and 2.33.

† Correlation = 1.00 represents no measurement error.

‡ Total number of cases in cohort. Calculated using Breslow and Day (15). Rounded to the nearest 100.

§ Assuming 5-year follow-up; 199.3 cases per 100,000 subjects per year. Rounded to the nearest 1,000.

To obtain the number of subjects required in the cohort, we calculate the proportion of new cases of disease, P, that we expect to see in the period of follow-up and divide the number of cases required, C, by P. For example, if the cohort comprised an equal number of subjects in the age ranges 50–54, 55–59, 60–64, 65–69, and 70–74 years, then one may calculate by using Surveillance, Epidemiology, and End Results Program incidence rates (16) that over a 5-year period an average of 199.3 new cases of colorectal cancer per 100,000 subjects per year would be observed. Hence, the proportion of cases, P, in the 5-year period is expected to be $5 \times 199.3 \times 10^{-5} = 0.009965$. Thus, assuming a correlation of 0.65, to reliably detect the small gradient hypothesis one would need a total of 10,400/0.009965 = 1,044,000 subjects. The numbers of subjects required, under different assumptions, for this colorectal cancer example are shown in table 5.

DISCUSSION

In this paper, we introduced model 1 to represent dietary measurement error and investigated the consequences relating to the distribution of true dietary intake in the population and the required sample size in a cohort study. The model is admittedly simplistic. Below we discuss some departures from the model which may occur in practice. Besides these, it must be acknowledged that there are many complexities in the analysis of cohort studies which we have not incorporated into our sample size calculations. For example, we have assumed that we are able to adjust adequately for all relevant confounders. If the confounders themselves are subject to measurement error, then bias in the estimated relative risks can result, with overestimation or underestimation occurring in different circumstances. It would be futile to try to include all such complexities in the sample size calculations, particularly since the main determinant of sample size, the relative risk gradient, is itself unknown. Nevertheless, a simple model such as model 1 can be useful to quantify the impact of measurement error on the required sample size.

Model 1 predicts that in the presence of dietary measurement error the true intakes will have a narrower distribution than those reported. Data from the Nurses' Health Study (2) and from the Women's Health Trial (12) comparing food questionnaires with the mean of several food records both show that the variances of intake ac-

cording to the food questionnaires were greater than the variances of those based on the mean of the food records. If we accept that the latter are likely to be a more accurate measure of true average intake, then these data are consistent with the prediction of model 1.

For nutrient intakes measured by weight or caloric content rather than as a percent of total calories, distributions tend to be skewed and are not normal. Logarithmic or square root transformations are often used to make the distributions of these intakes closer to normal, and model 1 might then apply to the transformed values of Y and X.

Several variants of model 1 relating reported intake Y to the true intake X are possible. For example, a more general model $Y = \alpha + \beta X + \epsilon$ may apply. In model 1, α is zero and β is equal to one, but different values of α and β may be envisioned. For example, Pietinen et al. (17) have noted a tendency for intake to be underreported on a questionnaire compared with the mean of several food records. Thus, α may in some cases take a negative value. However, such bias was not observed in the data on percent calories from fat in the Women's Health Trial (12).

A consequence of the more general model is that the ratio of standard deviation of X to Y equals ρ/β. When β equals one, this ratio becomes equal to ρ, a result mentioned earlier in the paper in connection with model 1. Thus, if model 1 is true, we should observe in validation studies that the ratio of standard deviations of food record intakes to questionnaire intakes is equal to the correlation between them. In fact, data on percent calories from fat from the Women's Health Trial (12) showed that in the "usual diet" group the ratio of standard deviations was 0.76, whereas the correlation was 0.67. This implies a value of β a little less than one ($0.67/0.76 = 0.88$). Similar estimates of β were seen for other macronutrient intakes. However, in the Nurses' Health Study (2), which employed a different questionnaire, estimated values of β for 16 macronutrient intakes were generally lower, ranging from 0.34 to 0.88 (mean, 0.62). This may reflect a tendency for those eating little of a nutrient to overreport their intake and for those eating large quantities to underreport the amount they consume, a pattern of reporting which would lead to a value of β less than one, an observation which has been referred to in the nutrition literature as the "flattened slope" effect. However, another factor to consider in these data is that the food record acts only as a surrogate for the truth. If model 1 were correct, error in the *food record* would tend to reduce below one the value of β in the regression between the questionnaire and the food record. Thus, an observed value for β of less than one could result from appreciable measurement error in the food record or, as noted above, from a departure from model 1, or both. Further analysis of food validation data may be able to disentangle these effects.

If model 1 were incorrect and the value of β were truly less than one, then the required sample size would be smaller than that based on model 1. This is because the standard deviation of the true intakes would be larger than ($1/\beta$ times) the value implied by model 1. When possible, accurate determination of β is advisable. However, often there is considerable uncertainty surrounding the value of β, and in this case it is prudent to adopt the model leading to the more conservative estimate of sample size, i.e., model 1. The possibility that β is greater than one, in which case even larger sample sizes would be required, is not supported by any data known to us, but cannot be entirely dismissed.

Another assumption of model 1 is that the error variance is independent of the true intake, i.e., the magnitude of reporting error does not change with the true intake. This may not always be the case. It might happen, for example, that the lower the true intake is, the smaller is the error. In fact, the Women's Health Trial data show a smaller estimate of the standard deviation of error (4.4 percent) in the intervention

group who ate a mean 21.1 percent calories from fat than in the control group (5.7 percent) who ate a mean 37.7 percent calories from fat. However, in this case the lower standard deviation in the intervention group may be due to the greater awareness of nutrition among these women and a consequently increased accuracy of dietary reporting. Further data, as they become available, will test the generality and applicability of model 1 to dietary measurement.

Besides the implied narrowing of the true intake distribution, measurement error also causes misclassification of subjects in the risk groups. Both phenomena lead to losses of power and increases in the required sample size of a cohort. Whereas Walker and Blettner (7) have focused only on the effects of misclassification, we have, by working with model 1, incorporated both effects into the sample size calculations.

Our sample sizes in table 5 seem particularly large, partly because of the effects of measurement error, but also because we have chosen relatively small gradients in relative risk as our hypotheses. We do this deliberately because even the smallest of these gradients may have important implications for public health. In table 6, we show the proportion of colorectal cancers which would be prevented were all members of the population to reduce their fat intake so as to move into the next lowest group (18), an average reduction of 7.5 percent calories from fat in the population. Under the large gradient hypothesis, this change in diet would prevent nearly 20 percent of colorectal cancer; were the small gradient hypothesis to be true, 7.5 percent of colorectal cancer would be prevented. This latter apparently small effect would amount to 11,000 fewer cases of colorectal cancer each year in the United States. Thus, the detection of such a small gradient in relative risk is worthwhile for common forms of cancer such as lung, breast, colorectal, and prostate cancers, and the conduct of a very large cohort study to discover such dietary associations deserves consideration.

An important question with regard to such a cohort study would be the level of bias which we might expect in assessing the effect of a nutrient, having corrected for known confounders and measurement error. There is a tradition that observed relative risks in epidemiologic studies need to be large to be convincing because of the possible biases which might operate. Early concern over bias was considerable, and it was suggested that relative risks of 2 or less were not to be regarded as conclusive evi-

TABLE 6

Preventable proportion of colorectal cancer were the population to shift their fat intake to the next lowest group

Percent calories from fat	Proportion with true intake in the range	Relative risk gradient		
		Small	Moderate	Large
<25.0	0.01	1.00	1.00	1.00
25.0–32.5	0.14	1.09	1.16	1.33
32.5–40.0	0.53	1.18	1.32	1.67
40.0–47.5	0.30	1.27	1.48	2.00
>47.5	0.02	1.36	1.64	2.33
Preventable proportion (%)*		7.5	11.7	18.9
Reduction in number of colorectal cancers per annum in the United States (current incidence, 147,000 per annum)		11,000	17,200	27,800

* If p_i is the proportion in the ith group (i = 1–5 in ascending order of fat intake) and r_i is the relative risk then: preventable proportion $= \sum_{i=2}^{5} p_i(r_i - r_{i-1}) / \sum_{i=1}^{5} p_i r_i$.

dence. Mantel and Haenszel (19), discussing case-control studies, are less conservative, suggesting 1.5 as the lower bound for conclusive evidence. However, in the same paper, they state that "A primary goal is to reach the same conclusions in a retrospective study as would have been obtained from a forward study, if one had been done" (19, p. 722). It is clear that, along with many others, Mantel and Haenszel considered that case-control studies are subject to a greater array of sources for potential bias than prospective cohort studies.

For cohort studies, it therefore would seem reasonable to interpret results somewhat less conservatively than for case-control studies. Table 4 shows that the small relative risk gradient leads to relative risks clearly too small to be convincing, the moderate gradient leads to relative risks which may still be a little too small, whereas the large gradient leads to relative risks which we feel would be convincing, if observed. Using our example of colorectal cancer it may therefore be appropriate to conduct a cohort study of size between that required for the large gradient (141,000) and the moderate gradient (401,000) to detect convincingly dietary risk factors with relative risks which are moderate but important from a public health perspective.

References

1. Block G. A review of validations of dietary assessment methods. Am J Epidemiol 1982;115:492–505.
2. Willett WC, Sampson L, Stampfer MJ, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. Am J Epidemiol 1985;122:51–65.
3. Byar DP, Freedman LS. Clinical trials in diet and cancer. Prev Med 1989;18:203–19.
4. Byers T. Diet and cancer. Any progress in the interim? Cancer 1988;62:1713–24.
5. Freudenheim JL, Marshall JR. The problem of profound mismeasurement and the power of epidemiological studies of diet and cancer. Nutr Cancer 1988;11:243–50.
6. Hebert JR, Miller DR. Methodologic considerations for investigating the diet-cancer link. Am J Clin Nutr 1988;47:1068–77.
7. Walker AM, Blettner M. Comparing imperfect measures of exposure. Am J Epidemiol 1985; 121:783–90.
8. Prentice RL, Pepe M, Self SC. Dietary fat and breast cancer: a quantitative assessment of the epidemiologic literature and discussion of methodologic issues. Cancer Res 1989;44:3147–56.
9. Ip C, Birt D, Rogers A, et al., eds. Dietary fat and cancer. Prog Clin Biol Res 1986;222:1–885.
10. Schoenborn CA, Marano M. Current estimates from the National Health Interview Survey: United States, 1987. Hyattsville, MD: National Center for Health Statistics, 1988 (Vital and health statistics, Vol. 10 (166)).
11. Liu K, Stamler J, Dyer A, et al. Statistical methods to assess and minimize the role of intra-individual variability in obscuring the relationship between dietary lipids and serum cholesterol. J Chronic Dis 1978;31:399–418.
12. Block G, Woods M, Sheppard L, et al. Comparison of a self-administered diet history questionnaire with multiple diet records. J Clin Epidemiol (in press).
13. Beaton GH, Milner J, Corey P, et al. Sources of variance in 24-hour dietary recall data: implications for nutrition study, design and interpretation. Am J Clin Nutr 1979;32:2546–59.
14. IMSL Library Users Manual. FORTRAN subroutines for mathematics and statistics. Ed. 9.2. Vol. 3. Chap M. Houston, TX: IMSL, Inc., 1984.
15. Breslow NE, Day NE. Statistical methods in cancer research: Vol II. The design and analysis of cohort studies. Lyon: IARC, 1987:285–7.
16. National Cancer Institute. Annual cancer statistics review. Washington, DC: US Department of Health and Human Services, 1988. pp. IIIB 30-1 (NIH publication no. 88-2789).
17. Pietinen P, Hartman AM, Haapa E, et al. Reproducibility and validity of dietary assessment instruments. II. A qualitative food frequency questionnaire. Am J Epidemiol 1988;128:667–76.
18. Wahrendorf J. An estimate of the proportion of colorectal and stomach cancers which might be prevented by certain changes in dietary habits. Int J Cancer 1987;40:625–8.
19. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959;22:719–48.

APPENDIX

Our method is described in steps A to F:

A. Assume we conduct a baseline survey in a sample which is representative of the cohort population. Let X = the true nutrient intake, and Y = the reported nutrient intake in this survey. Assume 1) Y is normally distributed with known mean μ_y and known standard deviation σ_y; 2) $Y = X + \epsilon$, where ϵ is normally distributed with mean 0 and is independent of X; 3) the correlation between Y and X is known, from validation studies of the baseline survey instrument, to be ρ. Then X is normally distributed, with mean $\mu_x = \mu_y$, and standard deviation $\sigma_x = \rho\sigma_y$. The latter result follows from $\rho = \dfrac{\mathrm{Cov}(X, Y)}{\sigma_x\sigma_y} = \dfrac{\sigma_x^2}{\sigma_x\sigma_y} = \dfrac{\sigma_x}{\sigma_y}$.

B. Assume that we question each member of the cohort study and that Z = the reported nutrient intake. Assume 1) Z is normally distributed with mean μ_z and standard deviation σ_z; 2) $Z = X + \delta$, where δ is normally distributed with mean 0 and is independent of X; 3) the correlation between Z and X is known, from validation studies of the cohort study instrument, to be τ. Since we already know from step A that X has mean μ_y and standard deviation $\rho\sigma_y$, and since the above assumptions imply that X has mean μ_z and standard deviation $\tau\sigma_z$, it follows that $\mu_z = \mu_y$ and $\sigma_z = \rho\sigma_y/\tau$.

In our calculations for the paper, we have assumed that $\tau = \rho$, so that $\sigma_z = \sigma_y$.

C. We divide the range of true intakes into n intervals: $(x_0, x_1), (x_1, x_2), \ldots, (x_{n-1}, x_n)$. We choose x_0 and x_n so that there is a negligible proportion of intakes beyond these limits. Let the relative risk for each group be $r_1, r_2, \ldots, r_n$; define the lowest intake group to be the reference so that $r_1 = 1$. These are the relative risk hypotheses shown in table 3 of the main text.

D. We divide the subjects into m groups, according to their reported intakes: $(z_0,z_1), (z_1,z_2), \ldots, (z_{m-1}, z_m)$. These groups could be chosen to be the same as $(x_0, x_1), \ldots, (x_{n-1}, x_n)$ or could be chosen according to another criteria, e.g., according to the quintiles of the reported intake distribution. Using the univariate normal distribution, we may calculate the probabilities Π_j of being in reported group j $(j = 1, \ldots, m)$. These are shown in table 1 of the main text. Using the bivariate normal distribution, we calculate the conditional probabilities:

$$p_{ij} = P[X \text{ is in } (x_{i-1}, x_i) \mid Z \text{ is in } (z_{j-1}, z_j)].$$

These are the probabilities shown in table 2 of the main text. In our calculations, we have chosen m and n to be equal to five and $(z_0, \ldots, z_m)$ to be equal to $(x_0, \ldots, x_n)$.

E. The relative risks $(q_1, \ldots, q_m)$ in the m groups based on reported intakes are calculated as:

$$q_j = \sum_{i=1}^{n} p_{ij}r_i \Big/ \sum_{i=1}^{n} p_{i1}r_i.$$

The denominator ensures that $q_1 = 1.0$, i.e., that group 1 acts as the reference group. These are the relative risks shown in table 4 of the main text.

F. The number of cases C required in the cohort is calculated using a method of Breslow and Day (15), assuming the analysis will employ a test for trend in relative risk over the m groups of reported intakes.

Let z_α be the z value corresponding to the significance test (e.g., for a two-sided 5 percent significance test $z_\alpha = 1.96$).

Let $z_{1-\beta}$ be the z value corresponding to the power of the test (e.g., for 90 percent power $z_{1-\beta} = 1.28$).

Let $w_1, \ldots, w_m$ represent the scores for the m groups $(z_0,z_1), \ldots, (z_{m-1},z_m)$, to be used in the trend test. For the calculations, we have assumed $w_j = j$ $(j = 1, \ldots, m)$.

Then

$$C = \frac{\left\{z_\alpha\left[\sum w_j^2\Pi_j - \left(\sum w_j\Pi_j\right)^2\right]^{1/2} + z_{1-\beta}\left[\sum q_j\Pi_j\left(w_j - \sum w_k\Pi_k\right)^2\right]^{1/2} \Big/ \left(\sum q_j\Pi_j\right)^{1/2}\right\}^2}{\left[\sum q_j\Pi_j\left(w_j - \sum w_k\Pi_k\right)\right]^2 \Big/ \sum q_j\Pi_j}$$

where summations are over the m reported intake groups. In this formula, terms in $C^{1/2}$ and lower have been ignored; the formula is a good approximation for large sample sizes, as occur in this paper. For problems in which the required numbers of cases are smaller, the more exact method given by Breslow and Day (15) should be used.

Walker and Blettner (7) implicitly assume that the distribution of X is known a priori. Essentially, they start out from stage B of our method and examine the effect on the power of different values of τ, varying τ while disregarding ρ. Our approach is therefore an extension of that of Walker and Blettner in which we also account for the imperfection of the dietary measurement in the baseline survey. By assuming $\tau = \rho$, we can, in one table (table 5), examine the joint effects of shrinkage of the true intake distribution and of misclassification. This could be done more generally assuming different values of ρ and τ, but would require several tables of sample size tabulated against τ (as in table 5), one for each value of ρ.